

# On international comparative studies on educational quality

Jan-Eric Gustafsson  
Gothenburg University



# Purposes

- Results from PISA, PIRLS, TIMSS and other international studies receive widespread attention in media, and these studies strongly influence discussions and decisions about educational policy. They also require major efforts and cost.
- How do these studies influence educational practice and educational research?
- In an attempt to answer these questions I will also discuss general methodological questions.

# Development of the international studies: phase one

- The International Association of Evaluation of Educational Achievement was founded in 1959 by a group of researchers in social science and education: *"The aim is to look at achievement against a wide background of school, home, student and societal factors in order to use the world as an educational laboratory so as to instruct policy makers at all levels about alternatives in educational organization and practice."*
- The First International Mathematics Study was conducted in 1964, and the Six Subject Survey in 1970-1971. A few studies were conducted in the 1980s, but in 1990 what I would like to describe as the first phase of the international studies came to an end.
- In phase one researchers were responsible for conceptualisation, design, analysis and reporting of the studies, as well as for fund raising.
- The studies required major efforts of everyone involved, but the scientific results were disappointing, because the cross-sectional survey design did not readily support causal inferences of the intended kind.

# Background of the international studies: phase two

- In 1990 the IEA established a new organization with a permanent headquarter in the Netherlands and a data-processing center in Hamburg.
- Shift from explanation to description, mainly serving the purpose of evaluation of educational quality as a basis for national discussions about educational policy.
- Limited researcher involvement, mainly with focus on methodological issues, and stronger involvement of national administrative and policy institutions.
- International reporting is primarily limited to description, while the task of finding explanations is left to participating countries and to the international research community to which the data is donated for secondary analysis.
- Dramatic increase in number of participating countries and frequency of studies, to which also the establishment of PISA (Program for International Student Assessment) by the OECD contributed greatly.

# Why is there a second phase?

- Increased focus on educational productivity and on output-driven modes of educational governance.
- New survey techniques became available in the 1980s, with advanced sampling methods and measurement technology based on item-response theory and matrix-sampling designs. The TIMSS 1995 study was the first study to adopt these techniques.

# Main questions

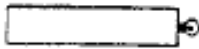
- The international studies provide an infrastructure of data that researchers can take advantage of, but they are not research studies in themselves. This raises two main questions:
  - Can we trust the descriptive results generated from these studies, i.e. are the international assessments of knowledge and skills reliable and valid?
  - Which kinds of conclusions do these studies support?

# Schoultz, Säljö & Wyndham (2001)

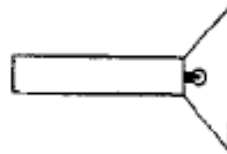
- Schoultz et al., (2001) questioned the validity of TIMSS 1995 because it is limited to the paper-and-pencil mode of assessment.
- They also argued that performance should be seen as produced through concrete communicative practice, rather than as a consequence of students' abilities and knowledge.
- They selected two items from the TIMSS 1995 study for scrutiny in an interview study comprising 25 Swedish grade 7 students.
- One was an optics item. It presented an illustration showing two flashlights, one with and one without a reflector, and the question was which of the two flashlights shines more light on a wall 5 meters away.

# Schoultz, Säljö & Wyndham (2001), cont

Jan and Lena each make a flashlight from identical batteries and bulbs. Lena's flashlight contains a reflector, while Jan's does not.



Jan's



Lena's

Which flashlight shines more light on a wall 5 meters away?

Explain your answer.

An open response was required, and to be scored correct the response had to include an explanation that argued that the reflector focused the light on the wall.

## Schoultz, Säljö & Wyndham (2001), cont

- In the Swedish Grade 7 TIMSS sample, only 39 % of the students answered the item correctly. However, in the interview study, 66 % of the students gave correct answers.
- The higher performance in the interview study was to a large extent due to the scaffolding provided by the interviewer in a Socratic dialogue.
- Schoultz et al. concluded that: “Knowing is in context and relative to circumstance. This would seem an important premise to keep in mind when discussing the outcomes of psychometric exercises.”
- This may seem as serious criticism not only of the validity of the TIMSS study, but also of results from paper and pencil tests generally. However, this study does not really address the issue of validity of the TIMSS assessment.

# Schoultz, Säljö & Wyndham (2001), cont

- Schoultz et al. start from the assumption that performance differences between different contexts are absolute, and that the higher performance in the interview situation is evidence of a higher level of knowledge and conceptual insight, i. e., evidence of higher student ability.
- In TIMSS, performance differences are seen as relative, because the observed performance level is interpreted as being determined not only by student ability but also by the difficulty of the item. According to this view, one possible interpretation is that the item is easier in the interview situation than in the paper and pencil situation.

# Assumptions and metaphors

- Is the absolute or relative view of performance differences correct? None, and both! Complex phenomena cannot be described unless we see them from a perspective. One way to capture different perspectives is to describe them in terms of metaphors.
- Sfard (1998) proposed that learning may be described in terms of either an acquisition metaphor or a participation metaphor.
  - The acquisition metaphor views knowledge as an acquired commodity, so learning is a process of acquisition with individual ownership of knowledge as a result.
  - The participation metaphor views learning as taking part in a collective and communicative process.
  - Both metaphors are limited, and we should not only choose one of them.

# Another metaphor: weather and climate

- Weather affects our daily lives, how we dress, what we do and talk about. We may adapt to weather but there is not much we can do about it. In the short run we can predict weather, but beyond a week or so weather is unpredictable.
- Climate is generalized weather. We experience weather, and through aggregating these experiences, we get a sense of climate. In a more precise manner scientists define climate as aggregate weather, using indicators such as mean temperature. Thus, climate is an abstraction.
- While weather is unpredictable and chaotic, climate and climate changes are stable phenomena, which we can understand theoretically and for which empirically based models may be constructed, that predict long-term development.
- In terms of this metaphor, large-scale survey studies are concerned with climate, while research which focuses on context-bound phenomena is concerned with weather.

# Is aggregation good or bad?

- Yanchar and Williams (2006) argued that: “... data aggregation and accompanying statistical tests often hide qualitative patterns and lead to excessively abstract or artificial conclusions ...; statistical indices are often used as facile substitutes for careful interpretation and human judgment ... patterns in aggregate data are erroneously used to make inferences about the structure of psychological processes in individuals ...”
- But the argument can also be turned around, and it can be argued that in order to see the general aspects (e. g., the climate) it is necessary to get rid of the specifics (e. g., the weather). Seen from this perspective, methods which conceal context-dependent variation have strengths, rather than disadvantages, when the purpose is to investigate general patterns and relations.
- Thus, aggregation may be both good and bad, depending upon the purpose of the research.

# Beyond the quantitative/qualitative dichotomy

- Much methodological debate starts from a dichotomy between quantitative and qualitative methods, and associated distinctions between objective/subjective, positivistic/hermeneutic, nomothetic/ideographic, and bad/good (e.g., Cohen, Manion, & Morrison, 2007).
- Ercikan and Roth (2006) argued that the quantitative and qualitative dichotomy is fallacious. They proposed that different forms of research should instead be put on a continuous scale that goes from the lived experience of people on one end (low-level inference) to idealized patterns of human experience on the other (high-level inference): “Knowledge derived through lower-level inference processes ... is characterized by contingency, particularity, being affected by the context, and concretization. Knowledge derived through higher-level inferences is characterized by standardization, universality, distance, and abstraction ....” (p. 20)

# Three level-of-inference dimensions

- Level-of-inference with respect to data:
  - High: data is generated through abstracting information over contexts and items.
  - Low: little or no abstraction over contexts.
- Level-of-inference with respect to generalization:
  - High: the aim is to generalize to a population.
  - Low: no aim to generalize to a population.
- Level-of-inference with respect to explanation:
  - High: the aim is to generate explanations, which take the form of causal relations among abstract constructs when high-level inference data is used
  - Low: the aim is to describe.

# Quality aspects of high-level inference data

- The purpose of high-level inference data is to capture abstractions, which span specific contexts and contents.
  - Is this possible and meaningful?
  - What criteria can be used to decide whether this is meaningful?
- The technology of measurement offers tools to do that, but quality cannot be judged in terms of a single number.
- Every step of the process of development and implementation of the large-scale assessments involves quality controls against explicitly defined criteria.
- The technological character of high-level inference data generation supports but does not guarantee the reliability and validity of the data.
- Does it work or not? By and large the patterns of results over time and over studies appear to be meaningful.
- According to Messick (1989), value implications and social consequences are important aspects of validity.

# Explanations in causal terms

- The original purpose of the international studies to find explanations has, to a large extent, been replaced with a descriptive aim.
- However, there still are official aims to generate explanations, and particularly so in the OECD studies. Furthermore, much of the secondary analyses done within the research community have explanatory aims.
- The descriptive results also require explanation, and if these are not supplied through research, they will be contributed by other stakeholders.

# Information technology availability and achievement

- Barber (2006) analyzed TIMSS 1999 data and concluded that information technologies mediate the effect of wealth on achievement. He also proposed that: “One practical implication is that even poor countries would be well advised to commit more of their budget to technological enrichment.”
- A report by OECD (2006) based on PISA data supports Barber’s conclusions that availability of computers is positively related to student achievement.
- Fuchs and Wößmann (2004) also analyzed PISA data. They too found that home availability of computers was positively correlated with student achievement. However, when they controlled for other variables the relationship changed.
  - Controlling for socio-economic background the relation vanished.
  - Controlling for school resources, the relation turned negative.
  - Controlling for institutional factors, the relation turned more strongly negative.
- Fuchs and Wößmann (2004) concluded that there are negative effects of availability of computers at home on student achievement.

# Why are the results contradictory?

- These studies try to make causal inferences from cross-sectional data. In such data relations between variables cannot be interpreted causally, because of:
  - Selection bias (or endogeneity effects, or reversed causality), which means that groups were not comparable to start with (e. g., kids with computers at home had higher achievement before there were any computers).
  - Omitted variables, which means that other variables which are correlated with the independent variable in focus, cause the relation (e. g., kids who have computers at home attend schools with qualified teachers).
- Unless we can get rid of these two sources of influence, we will make incorrect causal inferences.

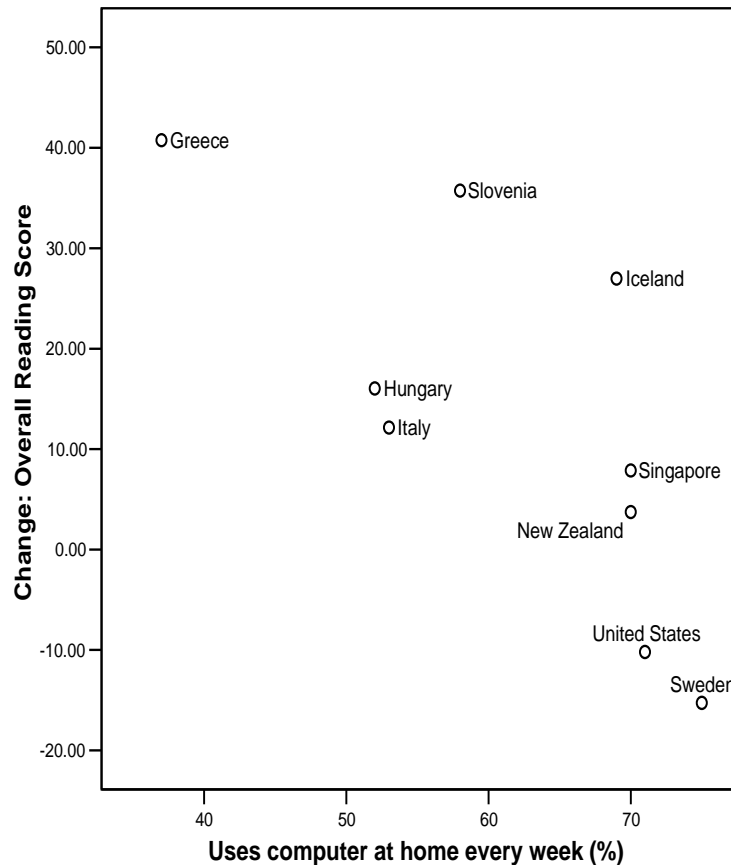
# How can the problems of selection bias and omitted variables be solved?

- Selection bias may be dealt with through analyzing data at a higher level of aggregation. For example, low-achieving students are often put in small classes. This will tend to cause class size to be positively correlated with achievement. But if there are multiple classes in each school, analyses at the school level will not be affected by such bias. Neither will analyses at the country level be affected by selection bias.
- Omitted variables may be dealt with through investigating change over time for fixed units, thereby making the units their own controls. This is the basic idea of longitudinal designs. With trend data this approach can be implemented at the country level.
- Thus, "fixed-country" analysis which relates change in independent variables to change in the dependent variable can be a way to reduce the problems of selection bias and omitted variables.

# An example of a country-level longitudinal analysis

- The IEA Ten-Year Trend Study (10YTS) gave estimates of change in reading scores during the 10-year period 1991 to 2001 for 9 countries (Greece, Hungary, Iceland, Italy, New Zealand, Singapore, Slovenia, Sweden and the USA).
- During these 10 years the home availability of computers had increased greatly in some countries and increased less in other countries.

# Relations between home availability of computers and change in reading score



# Correlations

- The correlation between frequency of computer use at home in 2001 and reading achievement was .08.
- The correlation between change in achievement between 1991 and 2001 and frequency of computer use at home in 2001 was  $-.73$  ( $p < .025$ ).
- The correlation between change in frequency of borrowing books at the library and change in reading achievement was  $.76$  ( $p < .008$ ).

# Conclusion

- These results suggest that increasing computer availability causes decreasing reading achievement, because computers detract students from reading activities.
- Thus, Barber (2006) and OECD (2006) are likely to have made invalid causal inferences.

# The importance of causal inference

- Many educational problems concern cause-and-effect issues.
- In the early 1980s most educational researchers turned away from issues of causality, for the same reasons that the IEA researchers no longer tried to use the world as an educational laboratory: they were disappointed and frustrated.
- Cronbach (1975) was overwhelmed by the complexities of educational phenomena, and the seeming impossibility of amassing empirical evidence in support of generalizations in the presence of multiple higher-order interactions between factors. Cronbach et al. (1980) published a volume on program evaluation that emphasized the context-dependence of approaches and results, rather than general principles and effects.
- In some other fields the focus on causal inference was not lost, and great progress has been made,

# Methods for causal inference from observational data

- *Instrumental variable estimation.* Controls for omitted variable bias if we can find a variable which is correlated with the independent variable that we try to determine the effect of, but which is uncorrelated with the omitted variables.
- *Propensity scores.* Controls for selection bias if we have information on background variables.
- *Regression discontinuity.* Those above a certain point on a continuous variable get the treatment, but not those below.
- *Longitudinal designs.* Analysis of fixed units controls for omitted variables.
- *Aggregation.* In some cases aggregation of data controls for selection bias.

# Conclusions

- The international studies generate an infra-structure of high-level inference data, which supports high-level generalization. The technology of measurement seems to work, even though there is room for improvement. Little attention has, however, been given to value implications and social consequences, which may be seen as important aspects of validity.
- Validity threats to causal inference are not always properly dealt with in analyses of the cross-sectional data, resulting in erroneous conclusions.
- The international studies have great potential for generating explanations in causal terms concerning interesting educational phenomena:
  - High-quality data from representative samples in many educational systems.
  - Data is longitudinal at the country-level, and through extensions of the basic design, it is sometimes possible to create a longitudinal design at the individual level as well.
  - New techniques for causal inference from observational data are available.
- If we take advantage of this potential, the international studies may prove to be exceptionally beneficial for improving practice and educational research.